THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Empirical Classification of Trajectory Data: An Opportunity for the Use of Machine Learning in Molecular Dynamics

Barry K. Carpenter,*[,†] Gregory S. Ezra,[‡] Stavros C. Farantos,[§] Zeb C. Kramer,[∥] and Stephen Wiggins[⊥]

[†]School of Chemistry, Cardiff University, Cardiff CF10 3AT, United Kingdom

[‡]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, United States

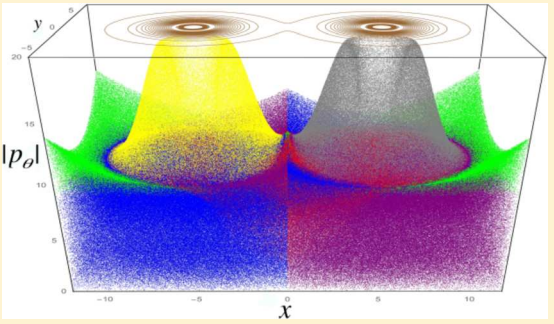
[§]Institute of Electronic Structure and Laser, Foundation for Research and Technology − Hellas, and Department of Chemistry, University of Crete, Iraklion 711 10, Greece

[∥]Department of Chemistry and Biochemistry, La Salle University, 1900 West Olney Avenue, Philadelphia, Pennsylvania 19141, United States

[⊥]School of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom

Supporting Information

ABSTRACT: Classical Hamiltonian trajectories are initiated at random points in phase space on a fixed energy shell of a model two degrees of freedom potential, consisting of two interacting minima in an otherwise flat energy plane of infinite extent. Below the energy of the plane, the dynamics are demonstrably chaotic. However, most of the work in this paper involves trajectories at a fixed energy that is 1% above that of the plane, in which regime the dynamics exhibit behavior characteristic of chaotic scattering. The trajectories are analyzed without reference to the potential, as if they had been generated in a typical direct molecular dynamics simulation. The questions addressed are whether one can recover useful information about the structures controlling the dynamics in phase space from the trajectory data alone, and whether, despite the at least partially chaotic nature of the dynamics, one can make statistically meaningful predictions of trajectory outcomes from initial conditions. It is found that key unstable periodic orbits, which can be identified on the analytical potential, appear by simple classification of the trajectories, and that the specific roles of these periodic orbits in controlling the dynamics are also readily discerned from the trajectory data alone. Two different approaches to predicting trajectory outcomes from initial conditions are evaluated, and it is shown that the more successful of them has ∼90% success. The results are compared with those from a simple neural network, which has higher predictive success (97%) but requires the information obtained from the "by-hand" analysis to achieve that level. Finally, the dynamics, which occur partly on the very flat region of the potential, show characteristics of the much-studied phenomenon called "roaming." On this potential, it is found that roaming trajectories are effectively "failed" periodic orbits and that angular momentum can be identified as a key controlling factor, despite the fact that it is not a strictly conserved quantity. It is also noteworthy that roaming on this potential occurs in the absence of a "roaming saddle," which has previously been hypothesized to be a necessary feature for roaming to occur.



1. INTRODUCTION

There is burgeoning interest in the application of machine learning (ML) methods to problems in computational chemistry,[1−21] with special attention having been paid to the fitting of potential energy surfaces (PES) for chemical reactions.[22−29] So far, there has been less attention paid to using ML techniques for analysis of chemical trajectory data,[8,29] although it appears that this could be a fruitful area of research. As a first step to undertaking such studies, we present here an analysis of trajectory data generated on a simple two degrees of freedom (2DoF) potential, which allows us to compare information derived from an empirical analysis of the trajectory data alone with that obtainable by a more conventional dynamical systems theory approach.

This paper is organized as follows. In the Method section we present the potential and describe the generation of the trajectory data from it. The Results and Discussion contains three subsections. In the first, we show how phase-space structures controlling the dynamics can be inferred from simple classifications of the trajectories, and how these are related to unstable periodic orbits that can be identified on the analytical potential. In the second subsection we present two different algorithms for predicting the outcomes of trajectories from their initial conditions alone. We compare them with each other

Special Issue: Benjamin Widom Festschrift

Received: September 1, 2017
Revised: October 2, 2017
Published: October 2, 2017

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

and with the predictive ability of a simple neural network. We also discuss how it can be possible to make outcome predictions in a formally chaotic dynamical system. In the final subsection we address the appearance of trajectories that show so-called roaming behavior[30−37] and discuss the possible implications for roaming in general. Finally, in the Conclusion we summarize the key findings and discuss the implications of the results for application of ML techniques to the analysis of trajectory data from classical molecular dynamics simulations.

2. METHOD

The potential studied in this work is related to the simple one-dimensional Morse function,[38] depicted in eq 1

$$V_{M1}(x) = D_e(1 - e^{-\sqrt{k/D_e}(x-r_e)})^2 \quad (1)$$

and graphically in Figure 1, which has long been used by chemists to represent a general dissociation of a diatomic

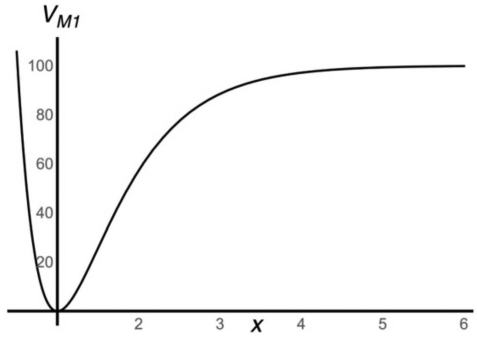


Figure 1. Graphical representation of eq 1, with parameter values D_e = 100, k = 200, r_e = 1.

molecule to two atoms. If one took the units of energy to be kcal/mol, and units of distance to be Å, the parameters for the Morse potential function shown in Figure 1 would roughly correspond to the vibration of a CH bond. A straightforward extension of this function comes from adding a second configuration-space dimension, as shown in eq 2 and Figure 2. The potential used in this work arises by allowing interaction between two identical two-dimensional Morse functions, as shown in eq 3 and Figure 3. This potential can be thought of as representative of the interaction of an atom with a fixed bond length, homonuclear diatomic molecule. The parameter b in eq

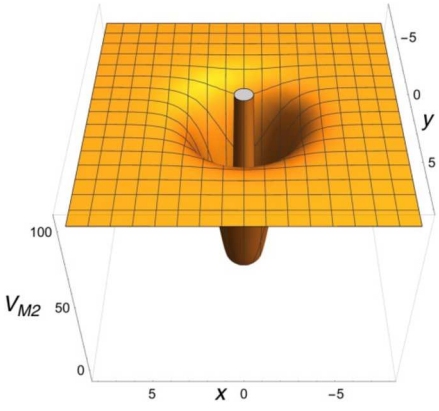


Figure 2. Graphical representation of eq 2, with parameter values D_e = 100, k = 200, r_e = 1.

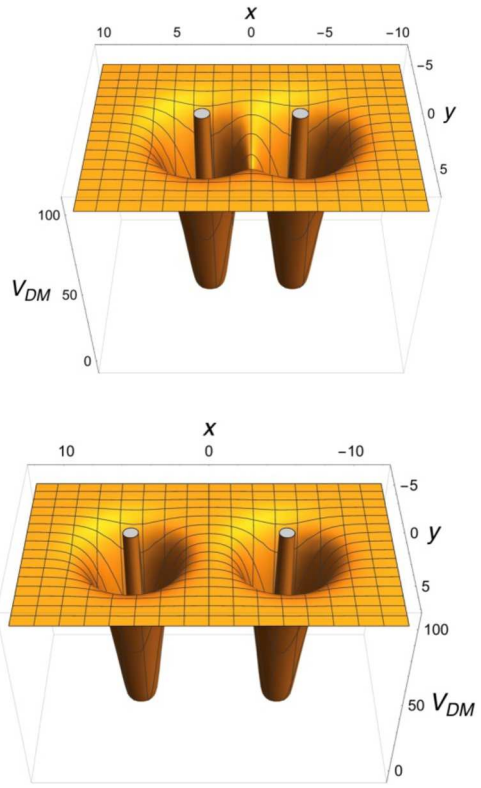


Figure 3. Graphical representations of eq 3, with parameter values D_e = 100, k = 200, r_e = 1. The upper panel has b = 3, and the lower panel has b = 5.

3 controls the distance between the two Morse functions and thereby the energy of the index one saddle created between them.

$$V_{M2}(x, y) = D_e(1 - e^{-\sqrt{k/D_e}(\sqrt{x^2+y^2}-r_e)})^2 \quad (2)$$

$$V_{DM}(x, y) = D_e(1 - e^{-\sqrt{k/D_e}(\sqrt{(x-b)^2+y^2}-r_e)})^2 + D_e(1 - e^{-\sqrt{k/D_e}(\sqrt{(x+b)^2+y^2}-r_e)})^2 - D_e \quad (3)$$

In general, the potentials studied here belong to the class of "cirques," characterized by one or more depressions in an otherwise flat plane.[39] We will refer to the particular subclass represented by eq 3 as double-Morse potentials. For all the work described in this paper, the value of b was set at 5. The Hamiltonian used for generation of the trajectories is shown in eq 4. It treats the masses of the diatomic molecule as effectively infinite, so that the center of mass of the system is at the origin. For all the results reported in this paper, the third body had unit mass (i.e., m = 1).

$$H(x, p_x, y, p_y) = \frac{p_x^2}{2m} + \frac{p_y^2}{2m} + V_{DM}(x, y) \quad (4)$$

The double-Morse potentials bear obvious similarity to that for the classic Euler three-body problem,[40] in which the attractive potential to each center varies as r^{-1}, where r is the distance of the third body from that center. However, the present potential has one important difference: the dynamics on the Euler potential are completely integrable whereas those on the present potential are not, as illustrated by the Poincaré surface of section in Figure 4. Figure 4 shows that the bound

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

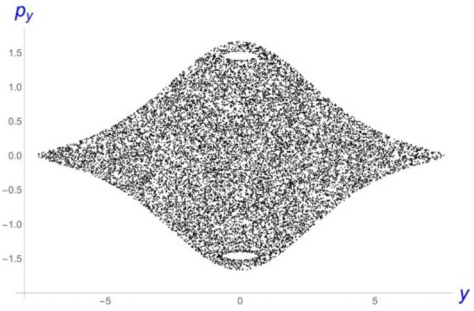


Figure 4. Poincaré surface of section for a trajectory initiated at $x = 0$, $y = 1$ and at a total energy of 99.995 on the double-Morse potential with parameter values $D_e = 100$, $k = 200$, $r_e = 1$, $b = 5$, $m = 1$. The trajectory was integrated for a total of 1.5×10^6 time units. The energy was kept slightly below that of the threshold (100 units) so that the trajectory remained bounded.

dynamics on the double-Morse potential are largely chaotic, although small islands due to a stable periodic orbit (PO) are apparent. Given that chaotic dynamics are usually associated with extreme sensitivity of outcome to initial conditions,[41] one might expect that it would be very difficult, at best, to predict outcomes of trajectories on this potential from their initial conditions. However, our interest is in trajectories with an energy slightly above that of the plane, and for these it does turn out to be possible to make outcome predictions with quite high success rates, as discussed below.

Trajectories on the double-Morse potential with $b = 5$ were initiated at random configuration-space coordinates in the rectangle $-15 \le x \le 15$, $-10 \le y \le 10$. The two Cartesian components of the momentum were selected randomly, but with the constraint that the total energy for each trajectory was 101 units. Trajectories were integrated using a velocity Verlet algorithm with a fixed time step of 10^{-4} units. One million trajectories were run, and the total energy was conserved to better than one part in 10^8 for each one. Because the total energy was chosen to be above the threshold energy of the flat part of the potential, every trajectory could in principle lead to dissociation of the third particle. Trajectories that had $|x| > 20$ or $|y| > 15$ were assumed to be on dissociative paths and were terminated. However, trajectories could also enter one or other minimum on the potential. Trajectories that achieved $r_A \le r_e$ or $r_B \le r_e$, where $r_A = \sqrt{(x+b)^2 + y^2}$ and $r_B = \sqrt{(x-b)^2 + y^2}$, were also terminated. From each starting point, a trajectory was integrated forward in time until it satisfied one of the three termination criteria and then backward in time until it satisfied one of those criteria. Trajectories, some of which are shown in Figure 5, were initiated at the blue dots and integrated forward and backward in time to the two red dots. Each trajectory consequently belonged to one of six connectivity classes, for which we use a two-letter designation, indicating the regions of the potential at the two termini of the trajectory. The initial conditions and connectivity classes for all 10^6 trajectories were recorded and, together, constituted the data set to be analyzed.

3. RESULTS AND DISCUSSION

Empirical Analysis of Trajectory Data. An obvious first step in the analysis of the trajectory data was merely to plot separately the configuration-space (e.g., $\{x,y\}$) coordinates of the initial conditions for each connectivity class. Unsophisticated though it may be, this exercise already revealed important dynamical features, as illustrated in Figure 6. Far from being randomly distributed across the sampled configuration space, for the 2DoF system studied here the initial conditions of the separate classes showed clear evidence of boundaries beyond which (inside in some cases and outside in others) no trajectories of the particular class could be found.

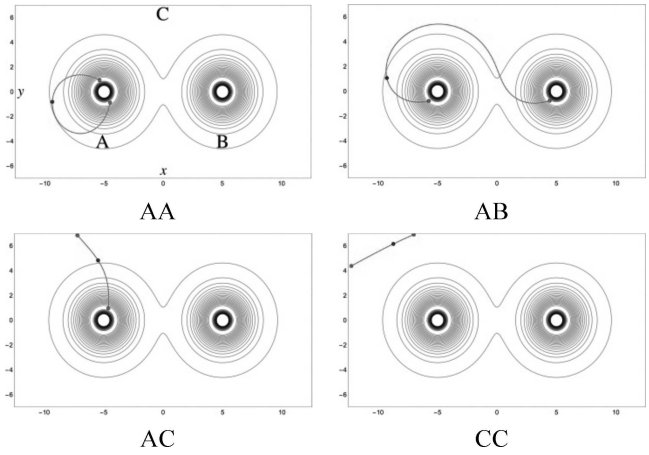


Figure 5. Representative trajectories corresponding to four of the six connectivity classes, superimposed on contours of the potential. As shown in the upper left panel, three regions of the potential, corresponding to the two minima and the flat plane, are labeled A, B, and C. Trajectories were classified by their termini in each of the regions. Trajectories terminating in the C region were, for the purposes of this figure, stopped before satisfying the criteria given in the text. Trajectories of classes BB and BC obviously exist, but are not shown because they are related by symmetry to the AA and AC classes, respectively.

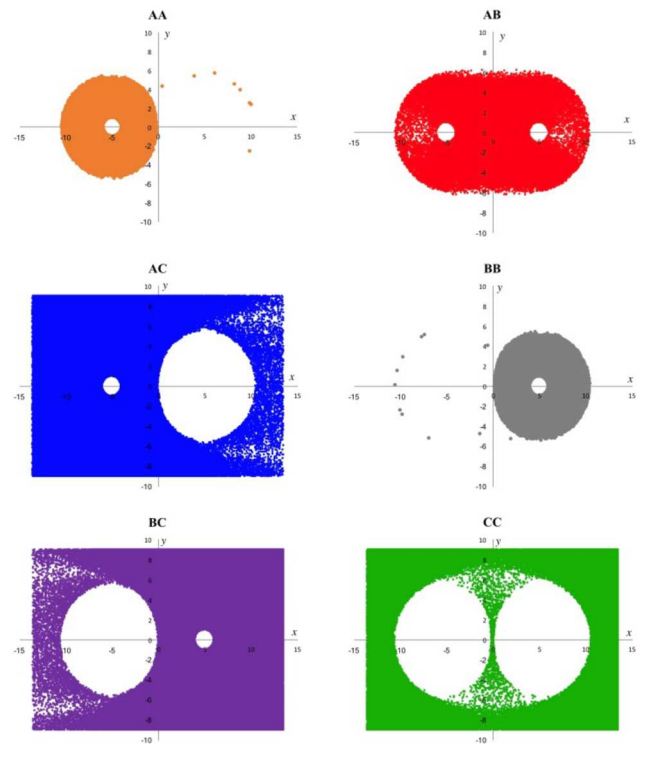


Figure 6. Distributions of initial conditions in configuration space for the six separate connectivity classes.

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

In addition to the apparent configuration-space boundaries between connectivity classes, the plots of initial conditions for the AA and BB classes in Figure 6 each show small numbers of points that are well separated from the principal groups, and seem possibly to be tracing out paths in configuration space. These points turn out to be initial conditions corresponding to so-called roaming trajectories.[30−37] Roaming is discussed in more detail below, in a subsection dedicated to the topic.

Clear relationships could be discerned between the individual data sets shown in Figure 6. The "non-roaming" AA and BB data points (i.e., those within the roughly circular principal data clusters of each class) were found to fit within the lacunae that are obvious in the AC, BC, and CC data sets. This is illustrated in Figure 7. Although Figure 6 makes it appear that the AA and

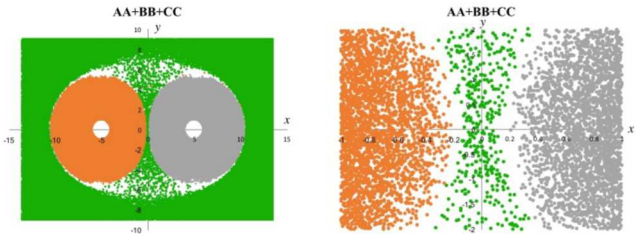


Figure 7. Relationship of the AA, BB, and CC data sets. The right-hand panel is a scale expansion near the origin, showing that there is no overlap among the different classes.

BB data points approach the origin of the plot more closely than they should if they did indeed fit within the lacunae, this is an artifact of the plots. It was necessary to plot the AA and BB data with a larger point size than the others in order to make the few roaming points visible. When all data are plotted with same point size, there is no overlap among AA, BB, and CC sets, as confirmed in the right-hand panel of Figure 7.

Further boundaries could be found when the momentum data of the initial conditions were included. The Cartesian components of the momentum did not prove to be particularly revealing, but their transformation into angular momentum values did. In particular, calculating the angular momentum with respect to an origin at the center of the A region of the potential (i.e., $\{-b,0\}$) for initial conditions with $x < 0$, and with respect to an origin at the center of the B region (i.e., $\{b,0\}$) for initial conditions with $x > 0$ revealed new boundaries in phase space, as illustrated in Figure 8. None of the AA or BB data points were found to have $|p_\theta| < 9.01$ units. This value appeared clearly in the phase space plots of all the data sets (see

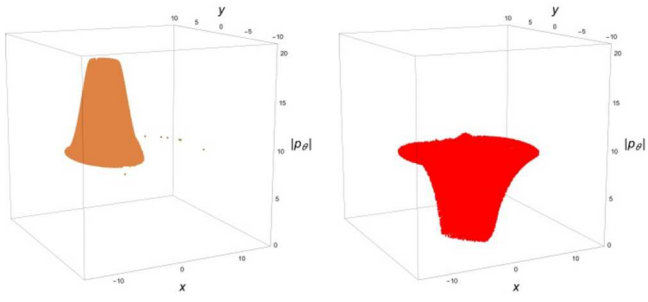


Figure 8. Phase space plots of initial conditions for the AA (left panel) and AB (right panel) data sets. Phase space plots for the other data sets are provided in the Supporting Information. Calculation of the angular momentum, p_θ, followed the prescription described in the text.

Supporting Information), although not always in the role of an inviolable boundary, as it seems to be for the AA and BB sets. The origin and significance of this value are described below.

In order to understand the boundaries between data sets that appear in the empirical analysis of the trajectory data, we choose here to use our prior knowledge about the underlying potential. This is obviously not a step that would be possible for trajectory data generated by direct molecular dynamics, but it serves to illustrate that the boundaries found in the data correspond to phase space objects that one can find by more conventional analysis of dynamics on an analytical potential. We hypothesize that the same would be true for other 2DoF potentials and that related effects might be seen in higher dimensional data.

On a 2DoF potential, boundaries between different types of trajectory behavior correspond to dividing surfaces (DS) built from unstable POs.[42] A detailed development of the construction of DSs for 2DoF systems using a certain class of unstable periodic orbits was given in a beautiful series of papers by Pollak, Pechukas, and Child in the 1970s and 1980s.[43−46] The resulting *periodic orbit dividing surfaces*, or PODS, possess many desirable features for determining the rate of crossing of trajectories of this DS. In particular, the DS has the "no-recrossing" property. Mathematically, this means that the Hamiltonian vector field is transverse to the DS. Another property of the DS constructed in this manner is that the flux across the DS is minimal, in the sense that perturbations to the DS lead to a larger flux. One might expect, then, that the boundaries depicted in Figures 6 and 7 would correspond to the configuration-space projections of such periodic orbits. This is indeed the case. Three types of unstable PO have been located (using the method described by Pollak, Pechukas, and Child) at an energy of 101 on the double-Morse potential with $b = 5$. They are illustrated in Figure 9.

The Type 1 and Type 2 POs come in counter-rotating pairs.[47] The Type 3 PO comes as a set of four, consisting of counter-rotating pairs around each of the PES wells. One can guess by inspection that the Type 1 PO forms the boundary around the AB data set depicted in Figure 6. This guess turns out to be correct. None of the initial conditions for AB trajectories could be found with configuration space coordinates outside the perimeter of the Type 1 PO; nor could initial conditions for the AA and BB trajectory classes. This information immediately gives one insight into the role of the Type 1 POs. The dividing surface built on the two Type 1 POs controls dissociation. Any point outside of these POs in configuration space will inevitably (for the selected total energy) yield a trajectory that will head off to infinity in positive and/or negative time integration. Such trajectories consequently cannot belong to the AA, AB, or BB connectivity classes. One might expect, given this analysis, that the initial conditions for CC class trajectories would all be found outside the Type 1 PO perimeter, but the data in Figure 6 show that this is clearly not the case. The reason is that there exist trajectories that can pass from infinity very close to the index one saddle and then back out to infinity, as illustrated in Figure 10. Note that the Type 3 POs do not quite pass through the index one saddle at the origin (see Figure 9). The CC trajectories of the kind just described must (for reasons outlined below) pass through the very narrow gap between the Type 3 POs surrounding the A and B wells. The probability of any trajectory having the initial conditions to do this is low, which is why the density of CC points inside the Type 1 PO

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

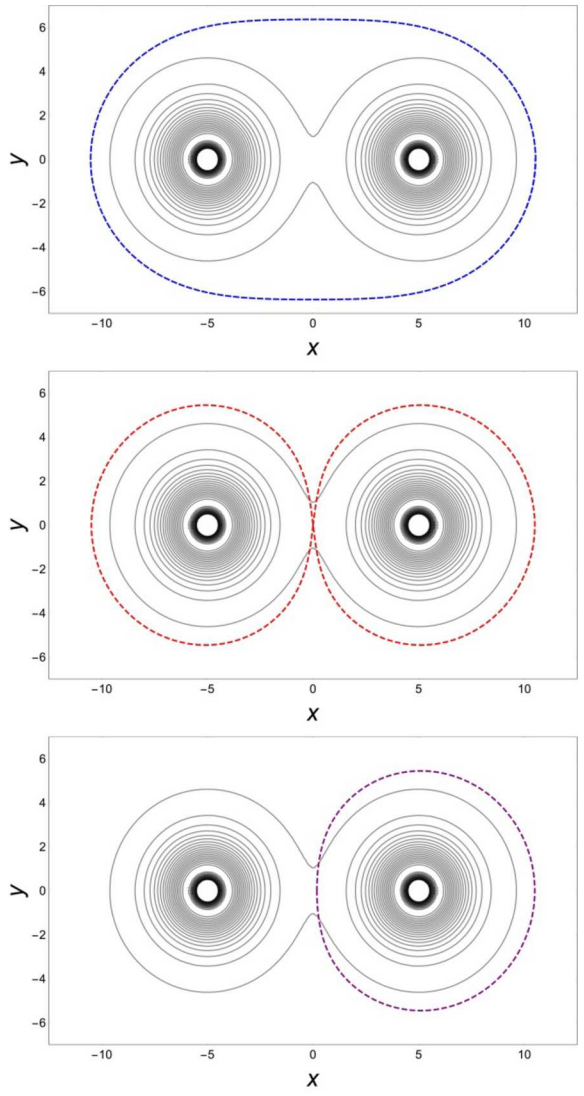


Figure 9. Configuration-space projections of three types of unstable periodic orbit found on the double-Morse potential with b = 5 at an energy of 101 units. They will be referred to as Type 1 (top), Type 2 (middle), and Type 3 (bottom). See text for further discussion.

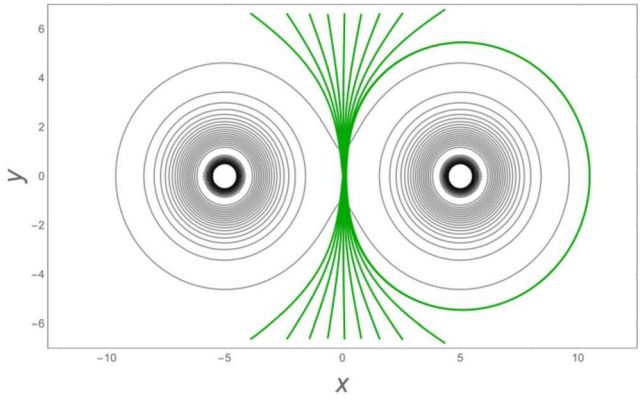


Figure 10. Trajectories initiated between the Type 2 and Type 3 POs are shown to be of the CC class, but passing close to the index one saddle at the origin of the potential.

perimeter is much lower than the density outside that perimeter (see Figure 6).

One can also guess by inspection that the Type 3 POs form the boundaries between the (non-roaming) AA and the CC data sets, and between the (non-roaming) BB and CC data sets. Again, this turns out to be correct. And, again, this information gives one insight into the roles of the dividing surfaces built on the Type 3 POs. It is apparent from Figure 6 that no AC class trajectory can ever pass inside the Type 3 POs around the B well. (It is perhaps worth pointing out that, although the trajectory data points are referred to as initial conditions in this paper, they actually correspond to arbitrary points along any of their respective trajectories, since integration backward in time from each point generates the true "initial" conditions for that trajectory.) Similarly, no BC class trajectory can ever pass inside the Type 3 POs around the A well. And no CC trajectory can ever pass inside any of the Type 3 POs. This latter point provides the reason that the few CC trajectories passing close to the index one saddle must squeeze between the Type 3 POs around each well. The reason that the Type 3 POs define these exclusion zones for some of the trajectories is that any trajectory (for the energy considered here) with configuration space coordinates inside one of the Type 3 POs will inevitably access the minimum in the potential for the PO in question. Hence, the Type 3 POs surrounding the A well all exclude trajectories belonging to connectivity classes that lack the letter A. A corresponding statement applies to the Type 3 POs surrounding the B well. Discussion of the role of the Type 2 PO is deferred until the subsection devoted to roaming.

The origin of the apparently special value of $|p_\theta| \approx 9.01$seen in the phase-space plots becomes apparent when one considers the angular momentum of each of the types of PO. As described earlier, the dynamics on this potential are not completely integrable, and so there are no conserved quantities beyond energy. Nonetheless, as Figure 11 shows, the angular momentum of the Type 1 PO is very nearly conserved, provided that one chooses the right origins for computing it.

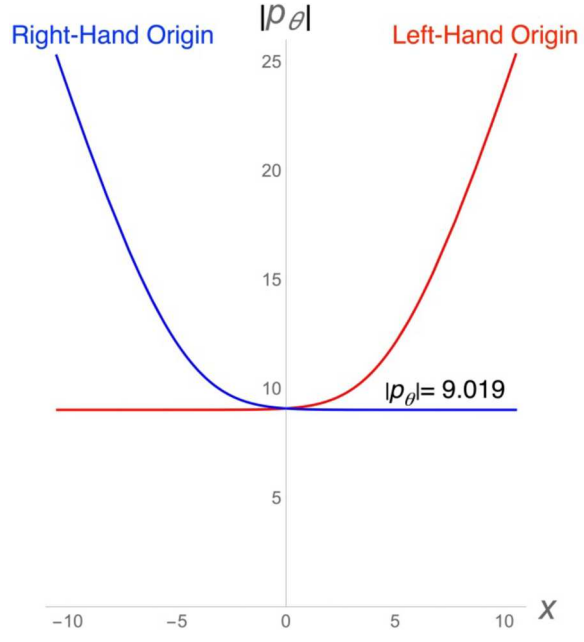


Figure 11. Absolute value of the angular momentum of the Type 1 PO on the double-Morse potential with b = 5 at an energy of 101. The red curve is plotted with respect to an origin at $\{-b,0\}$, and the blue curve, with respect to an origin at $\{b,0\}$.

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

One sees that the angular momentum of the Type 1 PO is very nearly constant all the time that the trajectory is in the half (defined by the sign of x) of the potential corresponding to the origin with which the angular momentum is calculated. The value of $|p_\theta|$ in the region $0.5 \leq |x| \leq 10.5$ is 9.019 ± 0.001 units. Similar results for the Type 2 and Type 3 POs are shown in Figure 12.

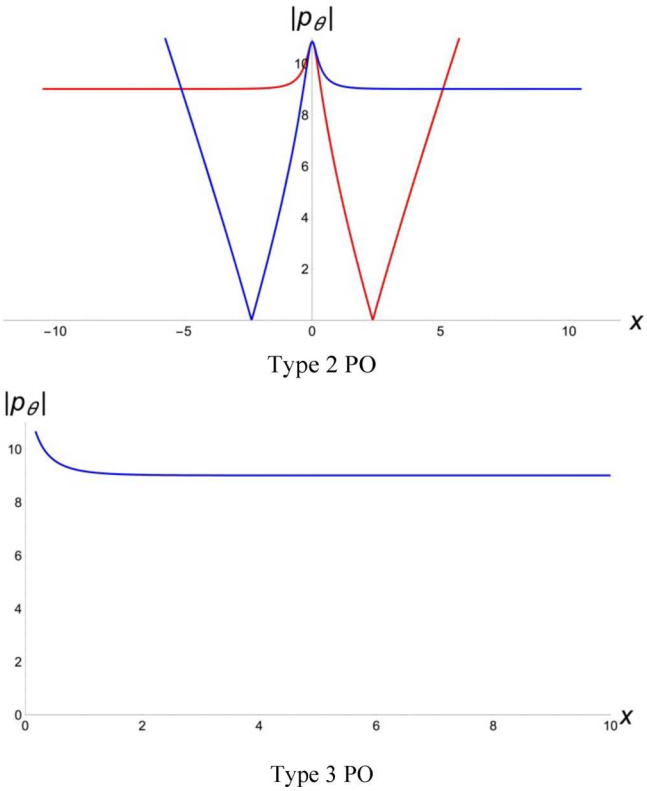


Figure 12. Absolute value of the angular momenta of the Types 2 and 3 POs on the double-Morse potential with $b = 5$ at an energy of 101. The red curve is plotted with respect to an origin at $\{-b,0\}$, and the blue curves, with respect to an origin at $\{b,0\}$.

The Types 2 and 3 POs clearly show larger variations in $|p_\theta|$ than the Type 1, as they approach $x = 0$, but nonetheless have effectively constant angular momentum for large parts of their excursions. For both POs, the value of $|p_\theta|$ in the region $2.0 \leq |x| \leq 10.5$ is 9.019 ± 0.003 units. Thus, in the regions where their angular momentum is close to constant, all three POs have very nearly the same value of $|p_\theta|$, and this value is very close to the boundary seen empirically in the phase-space plots of initial conditions. Not only do the three types of PO have very similar values of angular momentum, they also run very close to each other in configuration space for large parts of their excursions. An explanation for these observations is discussed in the Supporting Information.

In summary, then, the boundaries that one finds empirically between sets of initial conditions corresponding to the different connectivity classes correspond to the locations of key unstable periodic orbits in phase space.

Algorithms for Prediction of Outcomes for Trajectories from Their Initial Conditions. A common approach to simple, supervised tasks in ML is to assign a new input to one of a number of previously defined classes by calculating its distances from the centers of the data clusters for each class, specified in some chosen space.[48] The assignment is then made on the basis of the smallest distance. In the present case, the obvious analogous approach would be to compute coordinates in the phase space of the centers for the six clusters of initial conditions, corresponding to each connectivity class, and then to predict the outcome of any new trajectory by assigning it to the class to which its initial conditions are closest. Attempts to do this with the Cartesian coordinates and Cartesian components of the momentum were unsuccessful, but this is not very surprising, because the initial y coordinate, by itself, carries almost no information about the likely outcome of the trajectories on this potential, and because, as described above, the initial p_x and p_y values did not separate the different classes of trajectory very well in phase space. Consequently, some experimentation was undertaken to find combinations of phase space coordinates that might lead to better separation. The representation of the data leading to the best observed feature separation (although not proven to be the best possible) is summarized in eq 5 and Table 1. The square of the distance, R, for any point from the data center for each class was computed as shown in eq 5,

$$R^2 = w_1(r - r_c)^2 + w_2(x - x_c)^2 + w_3(|p_\theta| - |p_\theta|_c)^2 + w_4(|\tan^{-1}(p_x, p_y)| - |\tan^{-1}(p_x, p_y)|_c)^2 \quad (5)$$

where the w_i represent the four weights listed in Table 1, and $r = \sqrt{(x - \mathrm{sign}(x)b)^2 + y^2}$. No effort was made to normalize values of the four coordinates prior to the calculation, because the introduction of the weights for each coordinate allowed normalization constants to be combined with the weights. Initial guesses at the coordinates of each data center (terms with subscript c in eq 5) were made simply by computing averages for each parameter within a given connectivity class. However, because it was recognized that shapes of data clusters

Table 1. Coordinates of Data-Cluster Centers in Phase Space[a]

	AA	AB	AC	BB	BC	CC	weight
r_c	2.654	4.738	5.649	2.654	5.649	8.861	0.210
x_c	−5.062	0.000	−4.074	5.062	4.074	0.000	0.026
$\|p_\theta\|_c$	12.940	6.682	6.621	12.940	6.621	11.742	0.366
$\|\tan^{-1}(p_x,p_y)\|_c$	0.143	1.437	0.008	0.143	0.008	0.659	0.398
training set success	84.9	83.3	85.2	85.2	85.0	85.5	
test set success	85.7	82.3	85.6	84.6	84.0	84.8	

[a]Coordinate $r = \sqrt{(x - \mathrm{sign}(x)b)^2 + y^2}$. The angular momentum was calculated as described in the text.

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

may well be unsymmetrical, the final values of the coordinates and their weights were assigned on the basis of optimization by simulated annealing. The quantity optimized was not simply the overall success of the predictions, because that was found to lead to very uneven success rates between connectivity classes. Instead, the quantity optimized was $S/(1 + \sigma)$, where S is the overall success rate and σ is the standard deviation in success rates between each class. The success rates were the percentage of data points in a given connectivity class for which the predicted class was identical to the actual class. The training data set consisted of the original 10^6 trajectories described in the text. A test data set consisting of an additional 10^6 trajectories that had not been included in the calculation of fitting parameters was then evaluated using the optimized parameters, with results given in Table 1.

Although these results were encouraging, not least because they were obtained by a general approach that is commonly employed in ML algorithms,[48] it was recognized that not all the information presented in Section 2 of this paper had been utilized and that higher success rates might be achievable if it were. In particular, the localization of initial conditions for the six connectivity classes to different regions of configuration space implied that, if curves defining perimeters for these regions could be found, then outcomes could be predicted separately within each region. Such an approach would have two potential advantages. First, within some regions there would be only three possible outcomes instead of the six for the entire data set. (For example, outside of the curve corresponding to the Type 1 PO, only AC, BC, and CC classes occur.) Second, one could use different criteria to make predictions in different regions, potentially providing greater flexibility to the prediction algorithm.

In order to utilize this approach, it was necessary to define perimeter curves for the different configuration-space regions shown in Figure 13. For the present potential, we know that the perimeter curves correspond to the configuration-space projections of unstable periodic orbits, but in the case of trajectory data generated from direct dynamics simulations, one would not have that information. Consequently, an attempt was made to find perimeter curves simply from the trajectory data themselves. In the case of the curve corresponding to the Type 1 PO, the AB trajectory data were used. The curve corresponding to the left-hand Type 3 PO was generated from the AA, AC, and CC data, and that corresponding to the right-hand Type 3 PO, from BB, BC, and CC data. No attempt was made to find a curve corresponding to the Type 2 PO, because there were too few data points available to do so. In each case the largest area curve that excluded the appropriate data points, or the smallest area curve that contained the appropriate data points was sought. Each curve was described parametrically in terms of sine or cosine series in x and y. Detailed functional forms are given in the Supporting Information.

With the perimeter curves defined, each data point could be assigned to one of the regions of Figure 13. For the training data set, criteria were then sought that could best differentiate among the connectivity classes within an appropriate region. This was accomplished by plotting distribution functions of different coordinates or coordinate combinations for the separate connectivity classes within a region. Some representative examples are shown in Figures 14−17 for initial conditions found to be inside the perimeter that we know to represent the left-hand Type 3 PO.

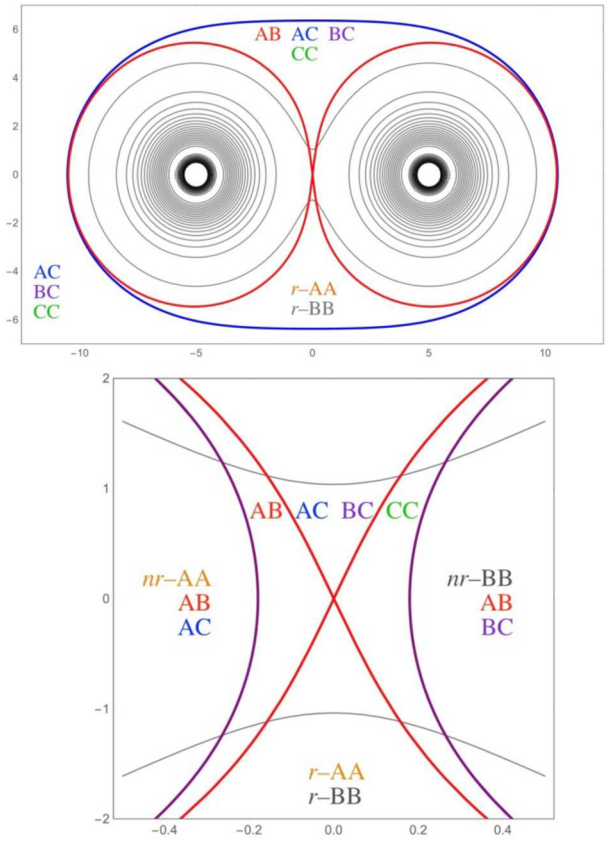


Figure 13. Localization of trajectory classes in configuration space. The lower panel is a scale expansion in the region of the index one saddle at the origin of the potential. The classes r−AA and r−BB are roaming trajectories, whereas nr−AA and nr−BB are non-roaming. The colored curves are periodic orbits, Type 1 in blue, Type 2 in red, and Type 3 in purple.

Best-fit functions were found for each of the distributions. When normalized to unit area, these functions provided empirical estimates of the probability that a trajectory in the chosen region and belonging to a particular connectivity class would have any given value of the parameter in question. Assignment of a new data point to a connectivity class then involved the following steps: first the point had to be assigned to one of the regions of configuration space shown in Figure 13. Next, the appropriate parameters for that region were computed from its initial conditions. The probability of each parameter value belonging to a given connectivity class was calculated from the distribution functions. Finally, all of the probabilities for the different parameters were multiplied to give an overall assignment probability. The connectivity class having the largest overall probability was chosen as the one to which the data point should be assigned.

The success rates by class for the training set and the test set are shown in Table 2.

The overall prediction success was 90.7% for the training set and 89.8% for the test set.

Preliminary work using neural network ML algorithms has shown significant further improvement in prediction success, with the best achieved so far being 97%. However, even with ML approaches, the success rate depends on the choice of phase space coordinates and momenta used, just as it did with analysis presented above. A combination of machine learning and physical insight is, at least in this example, more successful

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

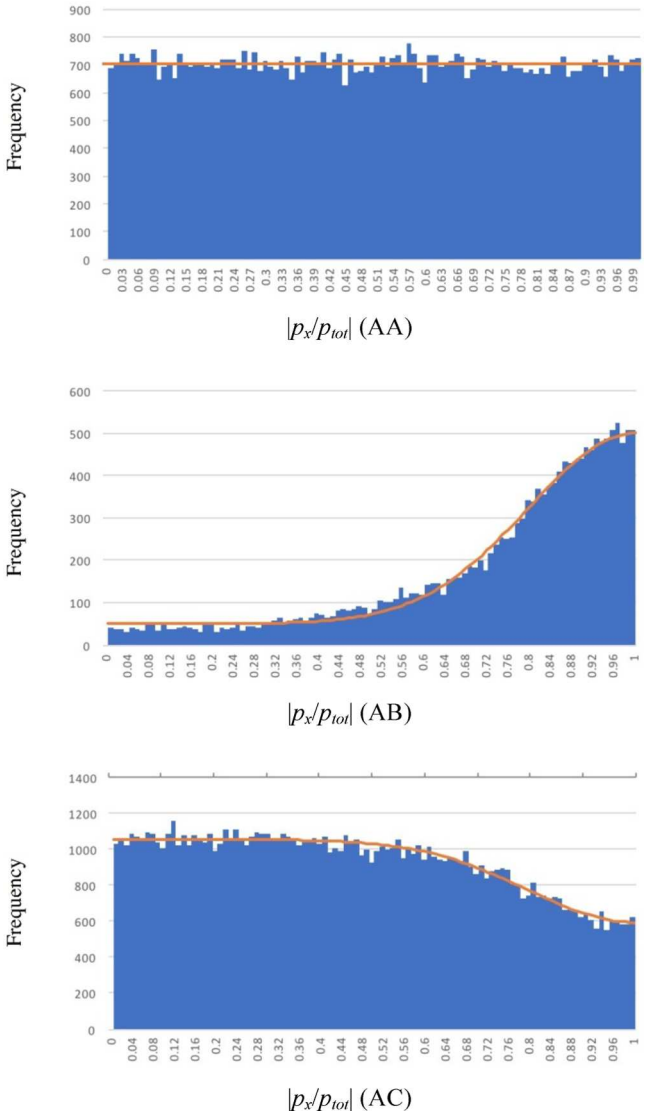


Figure 14. Distributions of the quantity $|p_x/p_{tot}|$ for the trajectories in the region described in the text. Blue bars represent data, and the orange lines represent best-fit functions.

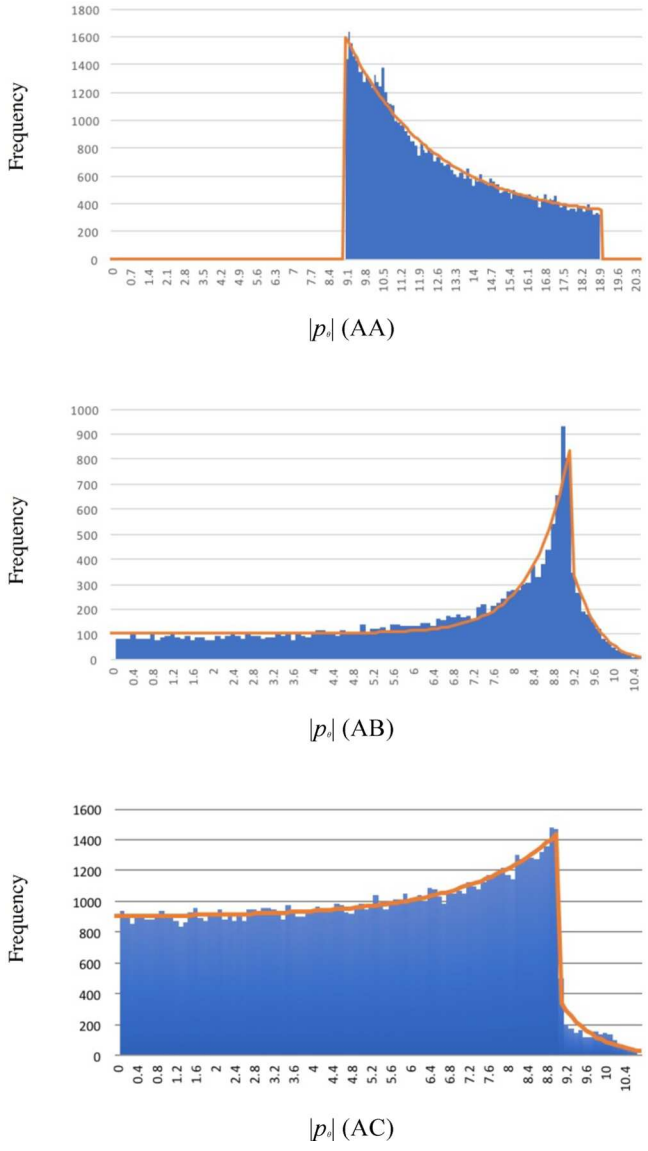


Figure 15. Distributions of the quantity $|p_\theta|$ for the trajectories in the region described in the text. Blue bars represent data, and the orange lines represent best-fit functions.

than either used alone. Details of the neural network are given in the Supporting Information.

An obvious question to be addressed is how it can be possible to obtain the results reported in this section, given the partially chaotic dynamics on the double-Morse potential. There are two probable components to the answer. The first concerns the energies involved in the generation of Figure 4 and in the generation of the data sets for the prediction exercises. The Poincaré surface of the section shown in Figure 4 came from tracking a single trajectory for 1.5×10^6 time units. In order to allow the trajectory to last that long in the vicinity of the minima, it was necessary to prevent it from heading off to infinity by choosing an energy slightly below the threshold value. However, the trajectories in the present data sets were generated with an energy 1% above that of the threshold. Hence, all of these trajectories were, in the limit of infinite time, unbound. Thus, the Poincaré surface of section shown in Figure 4, while serving to demonstrate that the dynamics on this potential are not completely integrable, is not representative of the dynamics at energies above the dissociation threshold. In fact, it is likely that at energies above threshold the present system exhibits the phenomenon of chaotic scattering, where finite-lifetime scattering trajectories coexist with a measure zero set of trapped (infinite lifetime) trajectories associated with a so-called "chaotic repellor."[49]

The second component of the answer concerns the durations of trajectories in the present study, which were typically <100 time units, because of the termination criteria described in Section 1. The termination criterion for trajectories entering the minima on the PES requires some thought. Given that total energy was conserved, it is clear that trajectories entering the minima would eventually have reemerged and, at some point, headed off to infinity. The justification for terminating them on first entry to a minimum is that this potential is supposed to be representative of that for a higher dimensional chemical system, in which there would be additional coordinates coupled to those considered here. Under such circumstances one expects the release of large amounts of kinetic energy, as occurs on entering a deep PES minimum, to be accompanied by at least

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

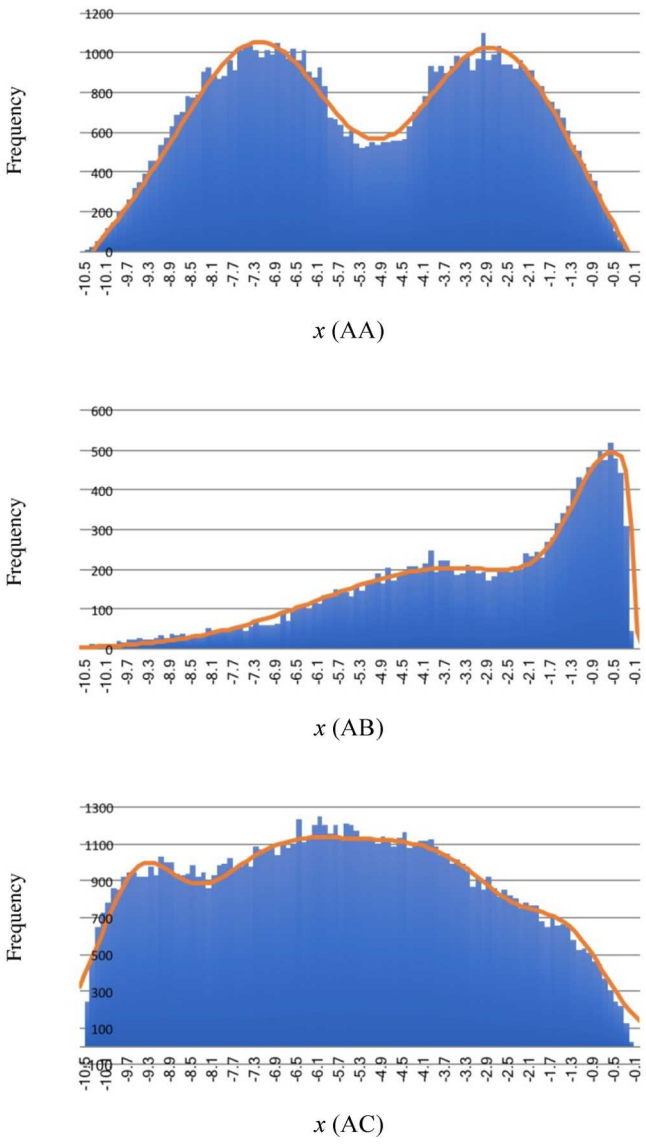


Figure 16. Distributions of x for the trajectories in the region described in the text. Blue bars represent data, and the orange lines represent best-fit functions.

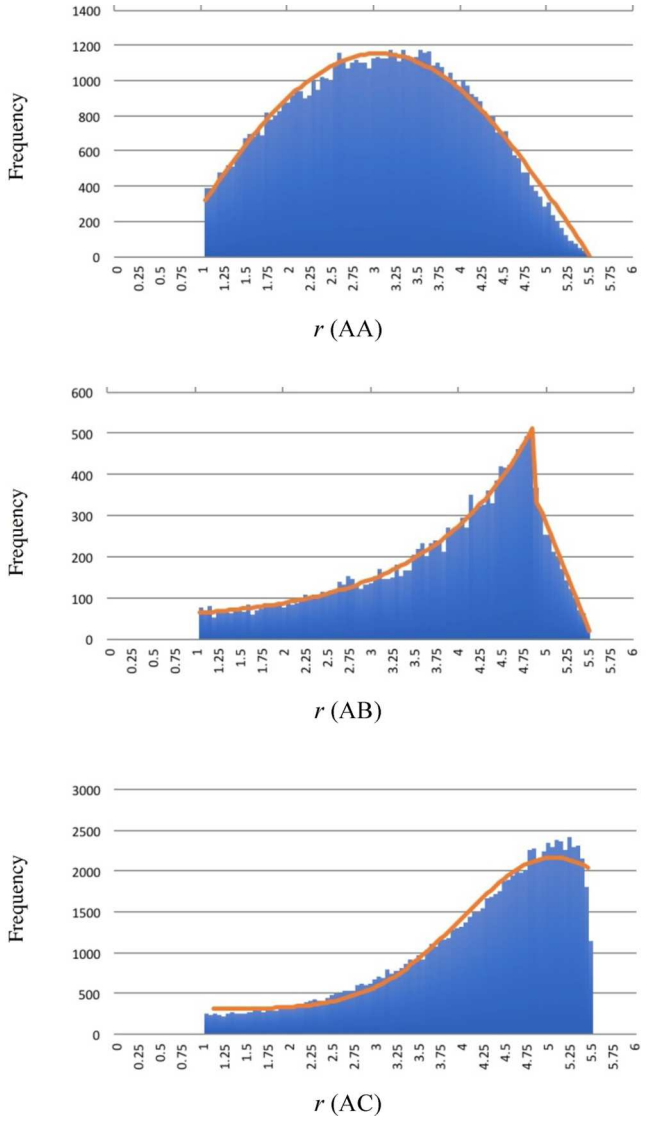


Figure 17. Distributions of r (distance from the center of the left-hand well) for the trajectories in the region described in the text. Blue bars represent data, and the orange lines represent best-fit functions.

some energy redistribution into the other coordinates. If the total energy is only slightly above the threshold, as it was in the present studies, then loss of even a small amount to other coordinates would effectively trap the trajectory in the minimum for an extended period. In general, one sees in classical molecular dynamics that there tends to be a relatively brief (often less than a picosecond) period of time during which reactive trajectories are on the highest energy parts of a PES. Even if the dynamics on the potential is intrinsically chaotic on the long (strictly speaking, infinite) time scale for which chaos is defined, during these short time intervals there may be a good deal more order to the dynamics than a chaotic description would imply. This phenomenon appears to lead to a more significant role for nonstatistical dynamical effects in chemical transformation than would have been expected if the dynamics were truly chaotic.[50]

Roaming Trajectories on the Double-Morse Potential. The AA and BB data plots in Figure 6 each revealed small numbers of points at values of x unexpected for their class—i.e. AA class with $x > 0$ and BB class with $x < 0$. These corresponded to points on "roaming" trajectories,[35] which start in one well and terminate in the same well, but do so by orbiting around the other well. Each of these data points had angular momentum close to the previously identified critical value of ~9.01. This observation allowed a search for more such points by oversampling in the angular momentum range 9.00–9.05. The results are shown in Figure 18, for the BB class. The initial conditions for the roaming trajectories were found to be bounded by the Type 1 and Type 2 POs, as illustrated in Figure 18. Plots of representative roaming BB trajectories are shown in Figure 19.

Table 2. Prediction Success Rates for the Separate Connectivity Classes, Using the Second Algorithm Described in the Text

	AA	AB	AC	BB	BC	CC
training set	98.6	89.6	89.0	98.8	89.9	90.1
test set	98.7	88.9	90.1	98.2	89.1	91.0

The occurrence of the roaming AA and BB trajectories is of some interest because in the extensive research on roaming it

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230–3241

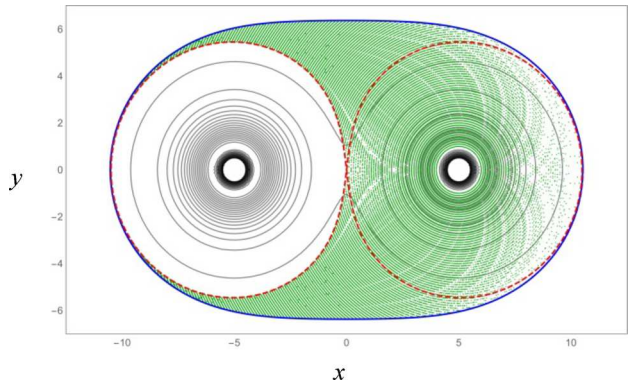


Figure 18. Configuration space plot of initial conditions for roaming BB trajectories. Type 1 (blue) and Type 2 (red) POs are superimposed.

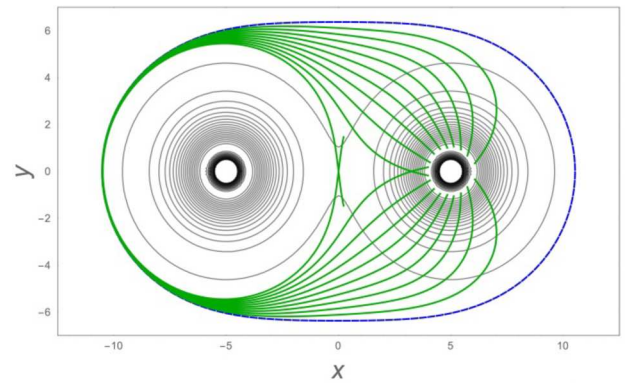


Figure 19. Roaming BB trajectories. The outer blue curve is the Type 1 PO. The innermost green trajectory is the left-hand half of the Type 2 PO.

has been hypothesized that the roaming phenomenon requires the existence of a special "roaming saddle" on the PES.[30−32,36,37] No such saddle exists on this potential.

One might characterize these trajectories as "unproductive roaming" since they begin and end in the same well. However, in prototypical chemical examples of roaming, such as in formaldehyde,[33] the corresponding trajectories would be productive. A hydrogen atom begins attached to the carbon atom of formaldehyde, roams around the oxygen, and then returns to the carbon. In the experimentally identifiable event called "roaming," the hydrogen atom does not then reattach to the carbon, but instead abstracts the second hydrogen. However, this last step, while important in making the roaming detectable, is not really the interesting part of the roaming event. It is the rotation of the hydrogen atom around the HCO fragment that has been the focus of attention and that is well reproduced in the roaming AA and BB trajectories here. The fundamental question in roaming has been why the roaming particle does not dissociate, despite having sufficient energy to do so.[34] The answer on the double-Morse potential considered here is that roaming occurs when trajectories happen to have initial conditions that bring them close in phase space to the unstable periodic orbits. Since all of the roaming trajectories have termini in one or the other well they can never exactly coincide with any of the periodic orbits shown in this paper, which never enter either well. However, it is possible to get asymptotically close to a periodic orbit, and the closer a trajectory gets, the longer it will spend roaming around on the PES. So, in short, roaming trajectories on this potential are simply "failed" POs. It may be that, on higher dimensional potentials, the roaming trajectories can similarly be considered "failed" NHIMs (normally hyperbolic invariant manifolds), which are the higher dimensional analogs of periodic orbits,[42] although that remains a conjecture for now. Additional analysis of the roaming trajectories on this potential is provided in the Supporting Information.

4. CONCLUSIONS

The principal conclusions from this work are as follows.

1. The different connectivity classes of trajectories on the double-Morse potential have initial conditions that are not randomly distributed either in configuration space or in phase space.
2. Because of the fact outlined in point 1, one can make probabilistic predictions of outcome for trajectories on this potential, given only their initial conditions. The highest success achieved to date for such predictions (97%) comes from a combination of physical insight and ML by neural network.
3. The dynamics on this potential are demonstrably chaotic below the threshold energy for dissociation. Nonetheless, this does not preclude high success in predicting outcomes for trajectories at energies above the dissociation threshold, even in the presence of chaotic scattering.
4. The phase-space structures that control the dynamics on this potential can be deduced from the trajectory data alone.
5. Roaming trajectories are found, despite the fact that the potential has no "roaming saddle."
6. The roaming trajectories are seen to be "failed periodic orbits," i.e. trajectories that come very close to the key unstable periodic orbits at some point in their transit.

Point 4 of the conclusions is arguably the most significant for future work. It holds out the possibility that empirical analysis of trajectory data, even when generated by direct dynamics techniques without benefit of an explicit potential energy surface, might yield insights into the phase space structures controlling the dynamics.

One way to view the results presented here is that they indicate how "reactivity boundaries" might be determined in an automated ML fashion.[51−55] The nature of reactivity boundaries for $N \geq 2$ DoF systems is the subject of ongoing research. For the high-dimensional data that would arise from typical molecular dynamics of polyatomic molecular systems, empirical analysis of reactivity boundaries would be very difficult to accomplish by hand, but may be amenable to machine learning techniques.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.7b08707.

Phase-space plots of initial conditions for separate trajectory connectivity classes; Relationships of periodic orbits on the double-Morse potential to trajectories on the 2D single-Morse potential; Details of perimeter curves and empirical probability distributions for the second prediction algorithm; Depiction of angular momentum changes for roaming trajectories; Description

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

of the neural network analysis of the trajectory data (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: carpenterb1@cardiff.ac.uk.

ORCID

Barry K. Carpenter: 0000-0002-5470-0278

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

B.K.C. thanks Dr. David Tew (University of Bristol) for helpful discussions. B.K.C. and S.W. acknowledge support from the Engineering and Physical Sciences Research Council of the UK, through Programme Grant EP/P021123/1. S.W. acknowledges the support of ONR Grant No. N00014-01-1-0769.

REFERENCES

(1) Bolis, G.; Dipace, L.; Fabrocini, F. A Machine Learning Approach To Computer-Aided Molecular Design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 617−628.

(2) Androulakis, I. P. New Approaches For Representing, Analyzing And Visualizing Complex Kinetic Transformations. *Comput. Chem. Eng.* **2006**, *31*, 41−50.

(3) Carrera, G. V. S. M.; Gupta, S.; Aires-de-Sousa, J. Machine Learning Of Chemical Reactivity From Databases Of Organic Reactions. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 419−429.

(4) Baldi, P.; Muller, K. R.; Schneider, G. Charting Chemical Space: Challenges And Opportunities For Artificial Intelligence And Machine Learning. *Mol. Inf.* **2011**, *30*, 751−752.

(5) Kayala, M. A.; Azencott, C. A.; Chen, J. H.; Baldi, P. Learning To Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209−2222.

(6) Qu, X. H.; Latino, D. A. R. S.; Aires-de-Sousa, J. A Big Data Approach To The Ultra-Fast Prediction Of DFT-Calculated Bond Energies. *J. Cheminf.* **2013**, *5*, 34−47.

(7) Fletcher, T. L.; Kandathil, S. M.; Popelier, P. L. A. The Prediction Of Atomic Kinetic Energies From Coordinates Of Surrounding Atoms Using Kriging Machine Learning. *Theor. Chem. Acc.* **2014**, *133*, 1499−1508.

(8) Li, Z. W.; Kermode, J. R.; De Vita, A. Molecular Dynamics With On-The-Fly Machine Learning Of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.

(9) Marcou, G.; de Sousa, J. A.; Latino, D. A. R. S.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. Expert System For Predicting Reaction Conditions: The Michael Reaction Case. *J. Chem. Inf. Model.* **2015**, *55*, 239−250.

(10) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic Spectra From TDDFT And Machine Learning In Chemical Space. *J. Chem. Phys.* **2015**, *143*, 084111.

(11) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning For Quantum Mechanical Properties Of Atoms In Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309−3313.

(12) Carr, S. F.; Garnett, R.; Lo, C. S. Accelerating The Search For Global Minima On Potential Energy Surfaces Using Machine Learning. *J. Chem. Phys.* **2016**, *145*, 154106.

(13) Himmetoglu, B. Tree Based Machine Learning Framework For Predicting Ground State Energies Of Molecules. *J. Chem. Phys.* **2016**, *145*, 134101.

(14) Li, L.; Baker, T. E.; White, S. R.; Burke, K. Pure Density Functional For Strong Correlation And The Thermodynamic Limit From Machine Learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *94*, 245129.

(15) Sadowski, P.; Fooshee, D.; Subrahmanya, N.; Baldi, P. Synergies Between Quantum Mechanics And Machine Learning In Reaction Prediction. *J. Chem. Inf. Model.* **2016**, *56*, 2125−2128.

(16) Abu El-Atta, A. H.; Hassanien, A. E. Two-Class Support Vector Machine With New Kernel Function Based On Paths Of Features For Predicting Chemical Activity. *Inf. Sci.* **2017**, *403*, 42−54.

(17) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction Of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(18) Galvelis, R.; Sugita, Y. Neural Network And Nearest Neighbor Algorithms For Enhancing Sampling Of Molecular Dynamics. *J. Chem. Theory Comput.* **2017**, *13*, 2489−2500.

(19) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning For Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.

(20) Pereira, F.; Xiao, K. X.; Latino, D. A. R. S.; Wu, C. C.; Zhang, Q. Y.; Aires-de-Sousa, J. Machine Learning Methods To Predict Density Functional Theory B3LYP Energies Of HOMO And LUMO Orbitals. *J. Chem. Inf. Model.* **2017**, *57*, 11−21.

(21) Yao, K.; Herr, J. E.; Parkhill, J. The Many-Body Expansion Combined With Neural Networks. *J. Chem. Phys.* **2017**, *146*, 014106.

(22) Handley, C. M.; Popelier, P. L. A. Potential Energy Surfaces Fitted By Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114*, 3371−3383.

(23) Bartok, A. P.; Gillan, M. J.; Manby, F. R.; Csanyi, G. Machine-Learning Approach For One- And Two-Body Corrections To Density Functional Theory: Applications To Molecular And Condensed Water. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 054104.

(24) Bartok, A. P.; Csanyi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.

(25) Rupp, M. Machine Learning For Quantum Mechanics In A Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.

(26) Behler, J. Perspective: Machine Learning Potentials For Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(27) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, And Outlook. *J. Phys. Chem. C* **2017**, *121*, 511−522.

(28) Kolb, B.; Marshall, P.; Zhao, B.; Jiang, B.; Guo, H. Representing Global Reactive Potential Energy Surfaces Using Gaussian Processes. *J. Phys. Chem. A* **2017**, *121*, 2552−2557.

(29) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning From MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726−741.

(30) Bowman, J. M. Roaming. *Mol. Phys.* **2014**, *112*, 2516−2528.

(31) Fu, B. N.; Bowman, J. M.; Xiao, H. Y.; Maeda, S.; Morokuma, K. Quasiclassical Trajectory Studies Of The Photodissociation Dynamics Of NO_3 From The D-0 And D-1 Potential Energy Surfaces. *J. Chem. Theory Comput.* **2013**, *9*, 893−900.

(32) Harding, L. B.; Georgievskii, Y.; Klippenstein, S. J. Roaming Radical Kinetics In The Decomposition Of Acetaldehyde. *J. Phys. Chem. A* **2010**, *114*, 765−777.

(33) Lahankar, S. A.; Chambreau, S. D.; Townsend, D.; Suits, F.; Farnum, J.; Zhang, X. B.; Bowman, J. M.; Suits, A. G. The Roaming Atom Pathway In Formaldehyde Decomposition. *J. Chem. Phys.* **2006**, *125*, 044303.

(34) Mauguiere, F. A. L.; Collins, P.; Kramer, Z. C.; Carpenter, B. K.; Ezra, G. S.; Farantos, S. C.; Wiggins, S. Roaming: A Phase Space Perspective. *Annu. Rev. Phys. Chem.* **2017**, *68*, 499−524.

(35) Townsend, D.; Lahankar, S. A.; Lee, S. K.; Chambreau, S. D.; Suits, A. G.; Zhang, X.; Rheinecker, J.; Harding, L. B.; Bowman, J. M. The Roaming Atom: Straying From The Reaction Path In Formaldehyde Decomposition. *Science* **2004**, *306*, 1158−1161.

(36) Tsai, P. Y.; Hung, K. C.; Li, H. K.; Lin, K. C. Photodissociation Of Propionaldehyde At 248 Nm: Roaming Pathway As An Increasingly Important Role In Large Aliphatic Aldehydes. *J. Phys. Chem. Lett.* **2014**, *5*, 190−195.

(37) Tsai, P. Y.; Li, H. K.; Kasai, T.; Lin, K. C. Roaming As The Dominant Mechanism For Molecular Products In The Photodissociation Of Large Aliphatic Aldehydes. *Phys. Chem. Chem. Phys.* **2015**, *17*, 23112−23120.

(38) Morse, P. M. Diatomic Molecules According To The Wave Mechanics. II. Vibrational Levels. *Phys. Rev.* **1929**, *34*, 57−64.

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241

(39) Quapp, W.; Kraka, E.; Cremer, D. Finding The Transition State Of Quasi-Barrierless Reactions By A Growing String Method For Newton Trajectories: Application To The Dissociation Of Methylenecyclopropene And Cyclopropane. *J. Phys. Chem. A* **2007**, *111*, 11287−11293.
(40) Euler, L. De Motu Corporis Ad Duo Centra Virium Fixa Attracti. *Nov. Comm. Acad. Sci. Petropolitanae* **1765**, *11*, 152−184.
(41) Lorenz, E. N. Deterministic Nonperiodic Flow. *J. Atmos. Sci.* **1963**, *20*, 130−141.
(42) Waalkens, H.; Wiggins, S. Geometrical Models Of The Phase Space Structures Governing Reaction Dynamics. *Regul Chaotic Dyn* **2010**, *15*, 1−39.
(43) Pechukas, P.; Pollak, E. Classical Transition State Theory Is Exact If The Transition State Is Unique. *J. Chem. Phys.* **1979**, *71*, 2062−2068.
(44) Pechukas, P.; Pollak, E. Trapped Trajectories At The Boundary Of Reactivity Bands In Molecular Collisions. *J. Chem. Phys.* **1977**, *67*, 5976−5977.
(45) Pollak, E.; Child, M. S.; Pechukas, P. Classical Transition State Theory: A Lower Bound To The Reaction Probability. *J. Chem. Phys.* **1980**, *72*, 1669−1678.
(46) Pollak, E.; Pechukas, P. Transition States, Trapped Trajectories, And Classical Bound States Embedded In The Continuum. *J. Chem. Phys.* **1978**, *69*, 1218−1226.
(47) Mauguière, F. A. L.; Collins, P.; Kramer, Z. C.; Carpenter, B. K.; Ezra, G. S.; Farantos, S. C.; Wiggins, S. Phase Space Barriers And Dividing Surfaces In The Absence Of Critical Points Of The Potential Energy: Application To Roaming In Ozone. *J. Chem. Phys.* **2016**, *144*, 054107.
(48) Samworth, R. J. Optimal Weighted Nearest Neighbour Classifiers. *Ann. Stat* **2012**, *40*, 2733−2763.
(49) Seoane, J. M.; Sanjuan, M. A. F. New Developments In Classical Chaotic Scattering. *Rep. Prog. Phys.* **2013**, *76*, 016001.
(50) Carpenter, B. K. Nonstatistical Dynamics In Thermal Reactions Of Polyatomic Molecules. *Annu. Rev. Phys. Chem.* **2005**, *56*, 57−89.
(51) Pechukas, P.; Pollak, E. Trapped Trajectories At Boundary Of Reactivity Bands In Molecular-Collisions. *J. Chem. Phys.* **1977**, *67*, 5976−5977.
(52) Tan, K. G.; Laidler, K. J.; Wright, J. S. Reactivity Bands In Atom-Molecule Collisions.III. Coplanar (H,H_2) Reaction. *J. Chem. Phys.* **1977**, *67*, 5883−5893.
(53) Wright, J. S. Reactivity Bands In Atom-Molecule Collisions 0.4. Coplanar And 3d Studies Of T+HT. *J. Chem. Phys.* **1978**, *69*, 720−724.
(54) Andrews, B. K.; Chesnavich, W. J. Boundary Trajectories In Collision-Induced Dissociation. *Chem. Phys. Lett.* **1984**, *104*, 24−27.
(55) Nagahata, Y.; Teramoto, H.; Li, C. B.; Kawai, S.; Komatsuzaki, T. Reactivity Boundaries For Chemical Reactions Associated With Higher-Index And Multiple Saddles. *Phys. Rev. E* **2013**, *88*, 042923.

DOI: 10.1021/acs.jpcb.7b08707
J. Phys. Chem. B 2018, 122, 3230−3241